

Bayesian Hypothesis Assessment in Two-arm Trials Using Relative Belief Ratios

Saman Muthukumarana and Michael Evans*

Abstract

This paper develops a Bayesian approach for assessing equivalence and non-inferiority hypotheses in two-arm trials using relative belief ratios. A relative belief ratio is a measure of statistical evidence and can indicate evidence either for or against a hypothesis. In addition to the relative belief ratio, we also compute a measure of the strength of this evidence as a calibration of the relative belief ratio. Furthermore, we make use of the relative belief ratio as a measure of evidence, to assess whether a given prior induces bias either for or against a hypothesis. Prior elicitation, model checking and checking for prior-data conflict procedures are developed to ensure that the choices of model and prior made are relevant to the specific application. We highlight the applicability of the approach and illustrate the proposed method by applying it to a data set obtained from a two-arm clinical trial.

Key words and phrases: equivalence, noninferiority, relative belief ratios, statistical evidence, bias induced from a prior, model checking and checking for prior-data conflict.

1 Introduction

Recently, hypothesis testing has been an active research topic in various types of two-arm clinical trials. As an example, a clinician may want to

*Saman Muthukumarana is Assistant Professor, Department of Statistics, University of Manitoba, Winnipeg, Manitoba R3T 2N2, Canada. Michael Evans is Professor, Department of Statistics, University of Toronto, Toronto, Ontario, M5S 3G3, Canada. The authors were partially supported by research grants from the Natural Sciences and Engineering Research Council of Canada.

demonstrate whether a new treatment is not worse than that of a reference treatment (also known as active control or standard treatment) by more than a specified margin [1]. This helps in assessing whether a less toxic, easier to administer, or less expensive treatment, is medically noninferior to a reference treatment. This kind of clinical trial, where the intention is to demonstrate that the new treatment is not inferior to the standard treatment by more than a small predefined margin δ , is known as a noninferiority trial. Here $\delta > 0$ is known as the prespecified clinically irrelevant or non-inferiority margin. Two-arm noninferiority trials of a new treatment and a well established reference treatment are an attractive option as in certain settings they avoid exposing patients to placebo situations. There has been a series of articles on this topic; see for example, special issues of *Statistics in Medicine* (Volume 47, Issue 1, 2005) and *Journal of Biopharmaceutical Statistics* (Volume 14, Number 2, 2004).

Sometimes in clinical trials, the goal is not to show that the new treatment is better, but rather equivalent to the reference treatment. This kind of clinical trial is known as an equivalence trial [2]. In an equivalence trial, the aim is to show that new treatment and the reference treatment have equal efficacy. For practical purposes, one can select a margin δ such that two treatments can be considered not to differ when their true difference lies in the interval of clinical equivalence $(-\delta, \delta)$. Note that this is different from testing a difference between two treatments which is a two-sided test known as superiority test in clinical literature. In this case $\delta > 0$ is called the clinically equivalence margin. Note that in either case, δ must be defined *a priori*.

There is a considerable literature on the related problems of hypothesis tests in clinical trials. As a simple frequentist method, one can use the standard t test for testing these hypotheses. At a higher level a generalized p-value approach may be applied using a generalized test function [3]. One can also perform tests using the ratio of the averages instead of the difference between the averages [4, 5]. Bayesian non-inferiority tests for proportions in two-arm trials with a binary primary endpoint [6] and normal means [7] are also considered in recent literature.

This paper considers a novel Bayesian approach for assessing a hypothesis in two-arm trials using relative belief ratios. A relative belief ratio for a hypothesized value of a parameter of interest is interpreted as the evidence that the hypothesized value is correct. We may obtain evidence either for or against the hypothesized value. Associated with a relative belief ratio is a measure of the strength of this evidence and this may be weak, strong or inconclusive. General inferences based on relative belief ratios are developed

in [8, 9, 10].

We discuss relative belief ratios and associated theory in Section 2 and illustrate this by application to the problem of interest. In Section 3 we consider the elicitation of the prior and how we can measure the suitability of the prior both with respect to its agreement with what the data say and with respect to measuring the bias a prior puts into the analysis. In Section 4, the approach is applied to a data set obtained from a two-arm clinical trial. We conclude with a short discussion in Section 5.

2 Inferences Based on Relative Belief Ratios

Suppose we have a statistical model, as given by a collection of densities $\{f_\theta : \theta \in \Theta\}$, and a prior π on Θ . After observing data x , the posterior distribution of θ is given by the density

$$\pi(\theta | x) = \frac{\pi(\theta)f_\theta(x)}{m(x)}$$

where $m(x) = \int_\Theta \pi(\theta)f_\theta(x) d\theta$. For an arbitrary parameter of interest $\psi = \Psi(\theta)$ we denote the prior and posterior densities of ψ by π_Ψ and $\pi_\Psi(\cdot | x)$, respectively. The relative belief ratio for a hypothesized value ψ_0 of ψ is defined by

$$RB_\Psi(\psi_0) = \frac{\pi_\Psi(\psi_0 | x)}{\pi_\Psi(\psi_0)}, \quad (1)$$

the ratio of the posterior to the prior at ψ_0 . As such, $RB_\Psi(\psi_0)$ is measuring how beliefs have changed that ψ_0 is the true value from *a priori* to *a posteriori*. Considering the case when the prior for ψ is discrete, we have that $RB_\Psi(\psi_0) > 1$ means that the data have lead to an increase in the probability that ψ_0 is correct, and so we have evidence in favor of ψ_0 , while $RB_\Psi(\psi_0) < 1$ means that the data have lead to a decrease in the probability that ψ_0 is correct, and so we have evidence against ψ_0 . As discussed in [10], this interpretation is also appropriate in the continuous case via a consideration of limits.

Clearly relative belief ratios are similar to Bayes factors, as they are both measuring change in belief, but the Bayes factor does this by comparing posterior to prior odds while the relative belief ratio compares posterior to prior probabilities and so is somewhat simpler. In fact, in certain circumstances the relative belief ratio and Bayes factor can be considered as equivalent but this is not always true. The full relationship between these quantities is discussed in [10].

One problem that both the relative belief ratio and the Bayes factor share as measures of evidence, is that it is not clear how they should be calibrated. Certainly the bigger $RB_{\Psi}(\psi_0)$ is than 1, the more evidence we have in favor of ψ_0 while the smaller $RB_{\Psi}(\psi_0)$ is than 1, the more evidence we have against ψ_0 . But what exactly does a value of $RB_{\Psi}(\psi_0) = 20$ mean? It would appear to be strong evidence in favor of ψ_0 because beliefs have increased by a factor of 20 after seeing the data. But what if other values of ψ had even larger increases? While calibrations of Bayes factors have been suggested [11, 12, 13] the proposed scales seem arbitrary and it is not at all clear that there is a universal scale on which Bayes factors or relative belief ratios can be calibrated.

A more useful calibration of (1) is given by

$$\Pi_{\Psi}(RB_{\Psi}(\psi) \leq RB_{\Psi}(\psi_0) \mid x) \quad (2)$$

which is the posterior probability that the true value of ψ has a relative belief ratio no greater than that of the hypothesized value ψ_0 . If we interpret $RB_{\Psi}(\psi_0)$ as the measure of the evidence that ψ_0 is the true value, we see that (2) is the posterior probability that the true value has evidence no greater than that for ψ_0 .

While (2) may look like a p-value, we see that it has a very different interpretation. For when $RB_{\Psi}(\psi_0) < 1$, so we have evidence against ψ_0 , then a small value for (2) indicates a large posterior probability that the true value has a relative belief ratio greater than $RB_{\Psi}(\psi_0)$ and so we have strong evidence against ψ_0 . If $RB_{\Psi}(\psi_0) > 1$, so we have evidence in favor of ψ_0 , then a large value for (2) indicates a small posterior probability that the true value has a relative belief ratio greater than $RB_{\Psi}(\psi_0)$ and so we have strong evidence in favor of ψ_0 . Notice that in the set $\{\psi : RB_{\Psi}(\psi) \leq RB_{\Psi}(\psi_0)\}$, the ‘best’ estimate of the true value is given by ψ_0 simply because the evidence for this value is the largest in this set. Various results have been established in [10] supporting both (1), as the measure of the evidence and (2), as the strength of that evidence.

As a measure of the strength of the evidence, (2) seems to work best when the posterior probabilities for all the possible values of ψ are all small or even 0 as in the continuous case. When some of these values have large posterior probabilities we can augment (2) as follows. If the prior π_{Ψ} corresponds to a discrete distribution with $\pi_{\Psi}(\psi_0) > 0$, we have that

$$\pi_{\Psi}(\psi_0 \mid x) \leq \Pi_{\Psi}(RB_{\Psi}(\psi) \leq RB_{\Psi}(\psi_0) \mid x) \leq RB_{\Psi}(\psi_0). \quad (3)$$

The right-hand inequality holds generally, see [10], while the left-hand inequality requires discreteness. Suppose $\pi_{\Psi}(\psi_0 \mid x)$ and (2) are both small

and notice that this happens whenever $RB_{\Psi}(\psi_0)$ is small. In this case we clearly have strong evidence against ψ_0 when $RB_{\Psi}(\psi_0) < 1$ and weak evidence for ψ_0 when $RB_{\Psi}(\psi_0) > 1$. Also, when $\pi_{\Psi}(\psi_0 | x)$ and (2) are both big, then we have only weak evidence against ψ_0 when $RB_{\Psi}(\psi_0) < 1$ and strong evidence for ψ_0 when $RB_{\Psi}(\psi_0) > 1$.

The other possibility is that the posterior probability $\pi_{\Psi}(\psi_0 | x)$ is small and (2) is big. If the prior probability $\pi_{\Psi}(\psi_0)$ is big and $RB_{\Psi}(\psi_0) < 1$, then this suggests that we have indeed obtained strong evidence against ψ_0 because (2) is big only because there are many other values of ψ for which there is evidence against ψ at least as strong as the evidence against ψ_0 . If, however $\pi_{\Psi}(\psi_0)$ is small and $RB_{\Psi}(\psi_0) < 1$, then we have weak evidence against ψ_0 because $\pi_{\Psi}(\psi_0 | x)$ is small due to $\pi_{\Psi}(\psi_0)$ being small. When $\pi_{\Psi}(\psi_0)$ is big and $RB_{\Psi}(\psi_0) > 1$, then we must have $\pi_{\Psi}(\psi_0 | x)$ is big as well, so no ambiguity arises, while when $\pi_{\Psi}(\psi_0)$ is small, then again $\pi_{\Psi}(\psi_0 | x)$ is small due to $\pi_{\Psi}(\psi_0)$ being small and so in both situations we have strong evidence in favor of ψ_0 via (2). So the only context where (2) might not suffice as a measure of the strength of the evidence given by $RB_{\Psi}(\psi_0)$, is when ψ has a discrete prior distribution with $\pi_{\Psi}(\psi_0)$ a non-negligible size, $\pi_{\Psi}(\psi_0 | x)$ small and (2) big. In general there is no harm in the discrete case in quoting (2), $\pi_{\Psi}(\psi_0 | x)$ and $\pi_{\Psi}(\psi_0)$, as part of the analysis of the strength of the evidence given by $RB_{\Psi}(\psi_0)$ and we recommend this.

There is another issue associated with using $RB_{\Psi}(\psi_0)$ to assess the evidence that ψ_0 is the true value. One of the key concerns with Bayesian inference methods is that the choice of the prior can bias the analysis in various ways. For example, in many problems Bayes factors and relative belief ratios can be made arbitrarily large by choosing the prior to be increasingly diffuse. This phenomenon is known as the Jeffreys-Lindley paradox because a diffuse prior is supposed to represent less information.

An approach to dealing with this paradox is discussed in [10]. Given that we accept that $RB_{\Psi}(\psi_0)$ is the evidence that ψ_0 is true, the solution is to measure *a priori* whether or not the chosen prior induces bias either for or against ψ_0 . To see how to do this we note first the Savage-Dickey result, see [14] and [10], which says that

$$RB_{\Psi}(\psi_0) = \frac{m(x | \psi_0)}{m(x)} \quad (4)$$

where

$$m(x | \psi_0) = \int_{\{\theta: \Psi(\theta) = \psi_0\}} \pi(\theta | \psi_0) f_{\theta}(x) d\theta$$

is the prior-predictive density of the data x given that $\Psi(\theta) = \psi_0$. Actually, it is easy to see that, if $T(x)$ is a minimal sufficient statistic for the full model, then $m(x | \psi_0)/m(x) = m_T(T(x) | \psi_0)/m_T(T(x))$ where m_T is the prior predictive density of T and $m_T(\cdot | \psi_0)$ is the prior predictive density of T given that $\Psi(\theta) = \psi_0$.

From (4) we can measure the bias in the evidence against ψ_0 by computing

$$M_T \left(\frac{m_T(t | \psi_0)}{m_T(t)} < 1 \mid \psi_0 \right) \quad (5)$$

as this is the prior probability that we will obtain evidence against ψ_0 when ψ_0 is true. So when (5) is large we have bias against ψ_0 . To measure the bias in favor of ψ_0 we choose values $\psi'_0 \neq \psi_0$ such that the difference between ψ_0 and ψ'_0 represents the smallest difference of practical importance. We then compute

$$M_T \left(\frac{m_T(t | \psi_0)}{m_T(t)} > 1 \mid \psi'_0 \right) \quad (6)$$

as this is the prior probability that we will obtain evidence in favor of ψ_0 when ψ_0 is false. Again, when (6) is large we have bias in favor of ψ_0 . Note that both (5) and (6) decrease with sample size and so, in design situations, they can be used to set sample size and so control bias.

When we are not able to control sample size, then (5) and (6) can be computed and used to qualify any conclusions we reach about whether ψ_0 is true or not. For example, if we have evidence against ψ_0 and (5) is large, this has to be taken with a ‘grain of salt’ as our choices have biased things this way. We draw a similar conclusion if we have evidence in favor of ψ_0 and (6) is large. Of course, these negative conclusions could also lead us to redo the analysis using a prior that does not induce such biases when this is possible, see Section 3.

A variety of other inferences can be derived from interpreting $RB_\Psi(\psi)$ as the evidence that ψ is the true value. For example, the best estimate of ψ is clearly the value for which the evidence is greatest, namely,

$$\psi_{LRSE}(x) = \arg \sup RB_\Psi(\psi),$$

and called the least relative surprise estimator in [8, 9, 10]. Associated with this is a γ -credible region

$$C_\gamma(x) = \{\psi : RB_\Psi(\psi) \geq c_\gamma(x)\} \quad (7)$$

where

$$c_\gamma(x) = \inf\{k : \Pi_\Psi(RB_\Psi(\psi) \geq k | x) \leq \gamma\}.$$

Notice that $\psi_{LRSE}(x) \in C_\gamma(x)$ for every $\gamma \in [0, 1]$ and so, for selected γ , we can take the size of $C_\gamma(x)$ as a measure of the accuracy of the estimate $\psi_{LRSE}(x)$. Given the interpretation of $RB_\Psi(\psi)$ as the evidence for ψ , we are forced to use the sets $C_\gamma(x)$ for our credible regions. For if ψ_1 is in such a region and $RB_\Psi(\psi_2) \geq RB_\Psi(\psi_1)$, then we must put ψ_2 into the region as well as we have at least as much evidence for ψ_2 as for ψ_1 .

In [8, 9, 10] various optimality properties are established for $\psi_{LRSE}(x)$ and the regions $C_\gamma(x)$ in the class of all Bayesian inferences. One notable property is that inferences based on the relative belief ratio are invariant under reparameterizations. This is not the case for Bayesian inferences based on losses, such as the posterior mean or mode and highest probability density regions.

We now consider the application of relative belief inferences to two-arm trials.

Example *Two-arm Trials.*

Let $x_E = (x_{E,1}, \dots, x_{E,n_E})$ denote the sample from the experimental treatment and $x_R = (x_{R,1}, \dots, x_{R,n_R})$ denote the sample from the reference treatment. We assume that these responses are mutually independent with $x_{E,i} \sim N(\mu_E, \sigma^2)$ and $x_{R,i} \sim N(\mu_R, \sigma^2)$ where $\mu_E, \mu_R \in R^1$ and $\sigma^2 > 0$ are all unknown. The information in the data is summarized by the minimal sufficient statistic $T(x_E, x_R) = (\bar{x}_E, \bar{x}_R, s^2)$ where $s^2 = [(n_E - 1)s_E^2 + (n_R - 1)s_R^2]/(n_E + n_R - 2)$ and the likelihood equals

$$\sigma^{-n_E - n_R} \exp\{-[n_E(\bar{x}_E - \mu_E)^2 + n_R(\bar{x}_R - \mu_R)^2 + (n_E + n_R - 2)s^2]/2\sigma^2\}.$$

We will use the prior for (μ_E, μ_R, σ^2) given by

$$\begin{aligned} \mu_E | \sigma^2 &\sim N(\mu_0, \tau_0^2 \sigma^2), \\ \mu_R | \sigma^2 &\sim N(\mu_0, \tau_0^2 \sigma^2), \\ 1/\sigma^2 &\sim \text{Gamma}(\alpha_0, \beta_0). \end{aligned} \tag{8}$$

We will discuss elicitation of the hyperparameters $\mu_0, \tau_0^2, \alpha_0$ and β_0 in Section 3. The posterior distribution of (μ_E, μ_R, σ^2) is then easily obtained and is given by

$$\begin{aligned} \mu_E | x_E, x_R, \sigma^2 &\sim N\left(\frac{n_E \bar{x}_E + \mu_0/\tau_0^2}{n_E + 1/\tau_0^2}, \frac{\sigma^2}{n_E + 1/\tau_0^2}\right), \\ \mu_R | x_E, x_R, \sigma^2 &\sim N\left(\frac{n_R \bar{x}_R + \mu_0/\tau_0^2}{n_R + 1/\tau_0^2}, \frac{\sigma^2}{n_R + 1/\tau_0^2}\right), \\ 1/\sigma^2 | x_E, x_R &\sim \text{Gamma}\left(\frac{n_E + n_R + 2\alpha_0}{2}, \frac{2\beta_0 + (n_E + n_R - 2)s^2}{2}\right). \end{aligned} \tag{9}$$

Note that it is simple to generate values from (8) and (9).

Now suppose we want to assess the hypothesis that the true value of $\mu_E - \mu_R$ satisfies $|\mu_E - \mu_R| < \delta$. So δ represents a practically meaningful difference between the means. If the difference is less than this quantity, then we do not distinguish between μ_E and μ_R but otherwise we do. It makes sense then that, if μ_E and μ_R do differ, we would want to know how many units of δ these means differed by. So for the ψ parameter of interest in this problem we will consider $\psi \in \mathbb{Z}$ where $\psi = i$ indicates that $\mu_E - \mu_R \in ((2i - 1)\delta, (2i + 1)\delta]$. So the hypothesis of interest corresponds to $H_0 : \psi = 0$.

To calculate the relative belief ratios for values of ψ we need the prior and posterior distributions of this parameter. These quantities are obtained by discretizing the prior and posterior distributions of $\mu_E - \mu_R$. We have that the marginal prior distribution of $\mu_E - \mu_R$ is given by

$$(\mu_E - \mu_R)/\tau_0 \sqrt{\frac{\beta_0}{\alpha_0}} \sim t_{2\alpha_0} \quad (10)$$

and the marginal posterior distribution of $\mu_E - \mu_R$ is given by

$$\{(\mu_E - \mu_R) - (\bar{x}_E - \bar{x}_R)\}/s_p \sqrt{1/n_E + 1/n_R} | x_E, x_R \sim t_\nu \quad (11)$$

where $\nu = n_E + n_R + 2\alpha_0 - 4$ and

$$s_p^2 = \frac{2\beta_0 + (n_E + n_R - 2)s^2}{n_E + n_R + 2\alpha_0 - 4}.$$

When ν is large, the posterior distribution is approximately normal, while it has heavy tails when ν is small.

So to assess H_0 the evidence is given by

$$RB_\Psi(0) = \frac{\Pi((-\delta, \delta] | x_E, x_R)}{\Pi((-\delta, \delta])} = \frac{\Pi((-\delta, \delta] | \bar{x}_E, \bar{x}_R, s^2)}{\Pi((-\delta, \delta])} \quad (12)$$

and the strength of the evidence is given by

$$\begin{aligned} & \Pi_\Psi(RB_\Psi(\psi) \leq RB_\Psi(\psi_0) | x_E, x_R) \\ &= \Pi(\cup_{RB_\Psi(i) \leq RB_\Psi(0)} ((2i - 1)\delta, (2i + 1)\delta] | x_E, x_R) \\ &= \sum_{i: RB_\Psi(i) \leq RB_\Psi(0)} \Pi(\{((2i - 1)\delta, (2i + 1)\delta] | x_E, x_R\}. \end{aligned} \quad (13)$$

Both (12) and (13) are easily evaluated using the exact distribution theory given for the prior and posterior distribution of $\mu_E - \mu_R$. For example, we can use the t distribution function routine in the R software package.

Suppose that we obtain $RB_\Psi(0) < 1$ and that (13) indicates that this is reasonably strong evidence against H_0 . From the tabulation of $RB_\Psi(i)$, that we computed as part of calculating (13), we easily obtain the optimal estimate of ψ , namely,

$$\psi_{LRSE}(x_E, x_R) = \arg \sup RB(i).$$

If $\psi_{LRSE}(x_E, x_R)$ is greater than 0, then we have a clear indication that the experimental treatment is better than the reference treatment. The accuracy of the estimate is assessed by computing the 0.95-relative belief region

$$C_{0.95}(x_E, x_R) = \{i : RB(i) \geq c_{0.95}(x_E, x_R)\}$$

and seeing how large it is. We can convert this into a region for $\mu_E - \mu_R$ via

$$C_{0.95}^*(x_E, x_R) = \bigcup_{i \in C_{0.95}(x_E, x_R)} ((2i - 1)\delta, (2i + 1)\delta].$$

To assess the bias in the prior we have to compute (5) and (6). From (10) and (11) we can evaluate

$$RB_\Psi(0) = \frac{\Pi((- \delta, \delta] | x_E, x_R)}{\Pi((- \delta, \delta])} = \frac{m_T(\bar{x}_E, \bar{x}_R, s^2 | - \delta < \mu_E - \mu_R \leq \delta)}{m_T(\bar{x}_E, \bar{x}_R, s^2)}$$

and note that $RB_\Psi(0)$ depends on the data only through $(\bar{x}_E - \bar{x}_R, s^2)$. We then need only simulate from the conditional prior predictive of $(\bar{x}_E - \bar{x}_R, s^2)$ given that ψ is the true value. Note that, given $(\mu_E - \mu_R, \sigma^2)$ then

$$\begin{aligned} \bar{x}_E - \bar{x}_R &\sim N(\mu_E - \mu_R, (1/n_E + 1/n_R)\sigma^2), \\ (n_E + n_R - 2)s^2/\sigma^2 &\sim \text{Chi-squared}(n_E + n_R - 2) \end{aligned} \quad (14)$$

and these quantities are independent.

We can compute (5) by the following simulation process:

1. set a counter $C = 0$,
2. generate σ^2 using (8),
3. generate $\mu_E - \mu_R$ from a $N(0, 2\tau_0^2\sigma^2)$ distribution conditioned to $-\delta < \mu_E - \mu_R \leq \delta$,
4. generate $(\bar{x}_E - \bar{x}_R, s^2)$ using (14),
5. compute $RB_\Psi(0)$ and add 1 to C if it is less than 1,

6. repeat 2-4 N times and record C/N as the estimate of (5).

Essentially the same simulation can be carried out to evaluate (6) with step 2 changing, as we condition on $(2i - 1)\delta < \mu_E - \mu_R \leq (2i + 1)\delta$ for say $i = 1$ or $i = -1$, and in step 4 we check if $RB_\Psi(0)$ is greater than 1.

The only slightly difficult part in this simulation is step 3 and for that we can use an inversion algorithm. For denoting the cdf and inverse cdf of a $N(0, 1)$ distribution by Φ and Φ^{-1} , respectively, we generate $\mu_E - \mu_R$ in step 3, when conditioning on $(2i - 1)\delta < \mu_E - \mu_R \leq (2i + 1)\delta$, by generating $u \sim U(0, 1)$ and putting

$$\mu_E - \mu_R = \Phi^{-1}(\Phi((2i - 1)\delta) + [\Phi((2i + 1)\delta) - \Phi((2i - 1)\delta)]u). \quad (15)$$

We can use routines in R for Φ and Φ^{-1} to evaluate (15).

3 Choosing and Checking the Ingredients

In any statistical analysis a statistician chooses a model that supposedly describes the generation of the data and, in a Bayesian analysis, also chooses a prior. As the analysis is typically highly dependent on these subjective choices, it is important that they be checked against what is typically objective, at least if it is collected correctly, namely, the data.

3.1 Checking the Model

For the model this entails asking if the observed data is surprising for every distribution in the model. If this is the case, then we conclude that there is a problem with the model and need to somehow modify this. While there are often many model checking procedures available, for the problem under study we will use the Shapiro-Wilks test based on the residuals from the model. We note that this check is, as it should be, completely independent of the choice of prior as we do not want to confound our considerations of the adequacy of the model and the prior.

3.2 Eliciting the Prior

Before discussing how we check the prior, we first consider the choice of the prior. For this we need only consider eliciting the prior for μ and σ^2 in a $N(\mu, \sigma^2)$ distribution. So we need to specify the hyperparameters $\mu_0, \tau_0^2, \alpha_0$ and β_0 . This is based on knowledge of the measurement process that leads to

the actual data and will typically require knowledge from someone familiar with making these kinds of measurements.

To elicit the prior for μ we specify an interval (m_1, m_2) that we are virtually certain (probability = 0.999) will contain this quantity. Of course we choose this as short as possible without being unrealistic. We then set $\mu_0 = (m_1 + m_2)/2$ and since

$$0.999 = \Phi\left(\frac{m_2 - \mu_0}{\tau_0 \sigma}\right) - \Phi\left(\frac{m_1 - \mu_0}{\tau_0 \sigma}\right) = 2\Phi\left(\frac{m_2 - m_1}{2\tau_0 \sigma}\right) - 1$$

we have that

$$\sigma^2 \leq ((m_2 - m_1)/2)^2 \{\Phi^{-1}((1 + 0.999)/2)\}^{-2} \tau_0^{-2}. \quad (16)$$

An interval that contains virtually all of the actual data measurements is given by $\mu \pm \sigma \Phi^{-1}((1 + 0.999)/2)$. Since this interval cannot be unrealistically too short or too long, we let s_1 and s_2 be lower and upper bounds on the half-length of the interval so that

$$s_1^2 \leq \sigma^2 \{\Phi^{-1}((1 + 0.999)/2)\}^2 \leq s_2^2. \quad (17)$$

Then equating the upper bound on σ^2 in (16) with the upper bound on σ^2 obtained from (17), we have

$$\tau_0^2 = ((m_2 - m_1)/2)^2 s_2^{-2}.$$

This determines the hyperparameters in the conditional prior for μ .

From (17) we have that

$$s_2^{-2} \{\Phi^{-1}((1 + 0.999)/2)\}^2 \leq 1/\sigma^2 \leq s_1^{-2} \{\Phi^{-1}((1 + 0.999)/2)\}^2. \quad (18)$$

Suppose again we want to determine the lower and upper bounds in (18) so that this interval contains $1/\sigma^2$ with virtual certainty. Therefore, letting $G(\alpha_0, \beta_0, \cdot)$ denote the $\text{Gamma}(\alpha_0, \beta_0)$ cdf we see that

$$\begin{aligned} G^{-1}(\alpha_0, \beta_0, (1 + 0.999)/2) &= s_1^{-2} \{\Phi^{-1}((1 + 0.999)/2)\}^2 \\ G^{-1}(\alpha_0, \beta_0, (1 - 0.999)/2) &= s_2^{-2} \{\Phi^{-1}((1 + 0.999)/2)\}^2 \end{aligned} \quad (19)$$

and we solve (19) for α_0 and β_0 by iteration. Noting that $G(\alpha_0, \beta_0, x) = G(\alpha_0, 1, \beta_0 x)$, and using the monotonicity of the cdf, leads to a simple iterative process.

So specifying the hyperparameters for (8) requires specifying an interval (m_1, m_2) that contains the true values of μ_E and μ_R with virtual certainty and also specifying the constants s_1, s_2 that specify lower and upper bounds on the length of any interval that will contain any measurement with virtual certainty. Of course, virtual certainty need not mean with probability 0.999 as some other large probability can be chosen. This value could be viewed as a conservative choice.

3.3 Checking the Prior

Checking the prior involves asking if the true value is a surprising value with respect to the prior. Methods for checking the prior in this sense are developed in [15, 17, 18]. Note that this is quite different than checking whether or not a prior induces bias as discussed in Section 2. A prior can avoid conflict with the data by being diffuse but at the same time induce bias into the analysis. Selecting a suitable prior involves balancing these considerations and tools have been developed for this.

The methods developed for checking the prior allows for all aspects of the prior to be checked simultaneously or for checking separate aspects of the prior in sequential fashion. The latter typically makes the most sense because, if we do detect prior-data conflict, then we will be better able to pinpoint where the problem lies.

The basic method for checking the prior involves computing, where T is the minimal sufficient statistic, the probability

$$M_T(m_T(t) \leq m_T(T(x))) \quad (20)$$

as this serves to locate the observed value $T(x)$ in its prior distribution. If (20) is small, then $T(x)$ lies in a region of low prior probability, such as a tail or anti-mode, which indicates a conflict. In the continuous case (20) is not invariant under 1-1 smooth transformations of the minimal sufficient statistic. Accordingly, (20) was modified in [16] to

$$M_T(m_T(t)/J_T(T^{-1}(t)) \leq m_T(T(x))/J_T(x)), \quad (21)$$

where $J_T(x) = (\det dT(x)(dT(x))^t)^{-1/2}$ and $dT(x)$ is the Jacobian matrix of T , to produce an invariant measure of prior-data conflict.

It is shown in [18] that, under quite general conditions, both (20) and (21) converge to $\Pi(\pi(\theta) \leq \pi(\theta_{true}))$, as the amount of data increases, where θ_{true} is the true value of the parameter. If $\Pi(\pi(\theta) \leq \pi(\theta_{true}))$ is small, then θ_{true} lies in a region of low prior probability which implies that the prior is

not appropriate. A logical approach to modifying a prior to avoid a conflict, when this is detected, is developed in [17].

For the prior given by (8) we will check this sequentially. First we will check the prior on σ^2 and, if no prior-data conflict is found, we then proceed to check the joint prior on (μ_E, μ_R) . For this we follow [15] which prescribes that the check on the prior for σ^2 is based on the prior predictive distribution of s^2 . Given σ^2 , we have that $V = (n_E + n_R - 2)s^2/\sigma^2 \sim \text{Chi-squared}(n_E + n_R - 2)$ and, as developed in the Appendix, a simple calculation gives that the prior predictive distribution is $V \sim ((n_E + n_R - 2)/\alpha_0)\beta_0 F((n_E + n_R - 2)/2, 2\alpha_0)$. Letting m_V denote the density of V and following [16, 18] an invariant p-value that checks the prior for σ^2 is given by

$$M_V(m_V(v)v^{1/2} \leq m_V((n_E + n_R - 2)s^2)(n_E + n_R - 2)^{1/2}s) \quad (22)$$

and this is easily computed via simulation.

If the prior for σ^2 has passed its check, then we can proceed to check the joint prior for (μ_E, μ_R) and this is based on the prior predictive distribution of $U = (\bar{x}_E, \bar{x}_R)$. Given σ^2 we have that

$$U \sim N_2 \left(\mu_0 \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \sigma^2 \begin{pmatrix} \tau_0^2 + 1/n_E & 0 \\ 0 & \tau_0^2 + 1/n_R \end{pmatrix} \right)$$

and an easy calculation presented in the Appendix gives that the prior predictive distribution is

$$U \sim t_{2\alpha_0} \left(2, \mu_0 \begin{pmatrix} 1 \\ 1 \end{pmatrix}, (\beta_0/\alpha_0) \begin{pmatrix} \tau_0^2 + 1/n_E & 0 \\ 0 & \tau_0^2 + 1/n_R \end{pmatrix} \right).$$

Denoting the prior predictive density of U by m_U , we check the joint prior for (μ_E, μ_R) via the p-value

$$M_U(m_U(u) \leq m_U(\bar{x}_E, \bar{x}_R)) \quad (23)$$

and this is easily computed using simulation. As discussed in [16, 18] this p-value is invariant because (\bar{x}_E, \bar{x}_R) is a linear function of the data.

4 Example

We illustrate the approach described in Sections 2 and 3 using a data set published in [2]. The data come from a comparative trial of moxonodin and captopril in the antihypertensive treatment of patients suffering from major depression. The response variable is the reduction of diastolic blood

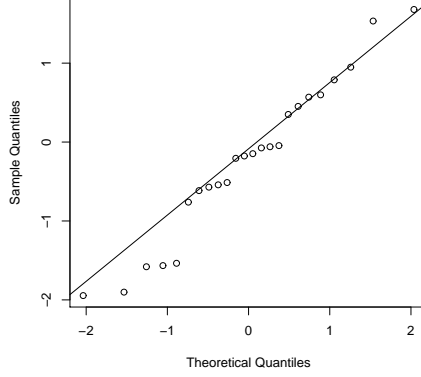


Figure 1: Quantile plots of experimental and reference treatment groups.

pressure, measured in millimeters of mercury (mm Hg), of patients suffering from a major depression under two drugs captopril (experimental) and moxonodin (reference). The data is given by

$$\begin{aligned} x_E &= (3.3, 17.7, 6.7, 11.1, -5.8, 6.9, 5.8, 3.0, 6.0, 3.5, 18.7, 9.6) \\ x_R &= (10.3, 11.3, 2.0, -6.1, 6.2, 6.8, 3.7, -3.3, -3.6, -3.5, 13.7, 12.6) \end{aligned}$$

so $n_E = n_R = 12$. The minimal sufficient statistic of the data is $T(x_E, x_R) = (\bar{x}_E, \bar{x}_R, s^2) = (7.21, 4.17, 46.79)$. We suppose that a practically meaningful difference in the means is given by $\delta = 0.5$ mm Hg.

Figure 1 gives the Q-Q plots of the residuals. The Shapiro-Wilks test for normality applied to the residuals gives a p-value of 0.51. This indicates that the data is not inconsistent with the normality assumption with constant variance.

For prior elicitation, we initially propose to reflect vague knowledge about the parameters by choosing a very diffuse prior. The values $m_1 = -100, m_2 = 100, s_1^2 = 5$ and $s_2^2 = 1000$ lead to the hyperparameter values

$$\mu_0 = 0, \tau_0^2 = 10, \alpha_0 \approx 2, \beta_0 \approx 5.$$

The bias in this prior is assessed by computing (4) and (5) with $\psi_0 = 0$. We found that

$$M_T(m_T(t|0)/m_T(t) < 1 \mid \psi = 0) = 0.07,$$

indicating that there is very little bias against ψ_0 . We see, however, that

$$M_T(m_T(t|0)/m_T(t) > 1 | \psi = 1) = 0.774$$

and this indicates that we have considerable bias in favor of ψ_0 . This is undoubtedly because we have chosen the prior to be too diffuse. Using such a prior will lead us to overstate the evidence in favor of the hypothesis or equivalently, understate the evidence against.

So to avoid the bias as much as possible, we chose different values for the hyperparameters. For this we set $m_1 = -20, m_2 = 20, s_1^2 = 10$ and $s_2^2 = 600$ which lead to the hyperparameter values

$$\mu_0 = 0, \tau_0^2 = 0.67, \alpha_0 \approx 1, \beta_0 \approx 8.$$

In this case we get

$$M_T(m_T(t|0)/m_T(t) < 1 | \psi = 0) = 0.49,$$

$$M_T(m_T(t|0)/m_T(t) > 1 | \psi = 1) = 0.40$$

indicating that there is some bias both for and against the hypothesis with this choice of prior. We cannot expect to be able to get both of these values to be low as this is controlled by sample size and we do not have a lot of data.

We checked this prior using (22) and (23). The value 0.15 was obtained for (22) and the value 0.22 obtained for (23). This indicates that there is no reason to be concerned about prior-data conflict.

Figure 2 contains a plot of the posterior and relative belief ratio for the continuous parameter. For the discretized parameter ψ we obtained $RB_\Psi(0) = 0.515$ with a strength of 0.19. As such we have only moderately strong evidence against the hypothesis of equivalence. Also

$$\psi_{LRSE}(x_E, x_R) = \arg \sup RB_\Psi(\psi) = 7$$

and the 0.95-relative belief region is given by

$$C_{0.95}^*(x_E, x_R) = (-0.5, 13.5].$$

The length of this interval indicates a fair degree of uncertainty about the true value but we do have reasonable evidence that the treatments are not equivalent.

Note that $C_{0.95}^*(x_E, x_R)$ includes the value 0 but this is not a contradiction with the fact that we have evidence against $\psi = 0$. These credible

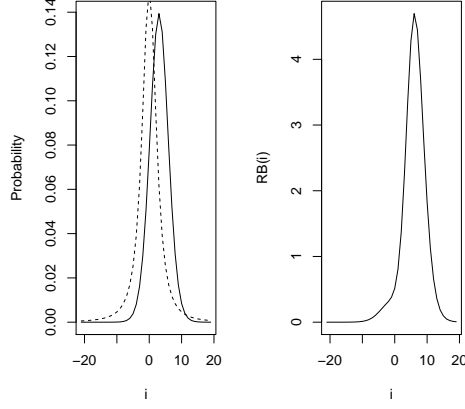


Figure 2: Plots of prior (---) and posterior (—) densities and of the relative belief ratio (—) of ψ .

regions do not work like confidence intervals do with p-values. It is only the length of $C_{0.95}^*(x_E, x_R)$ that is relevant as a measure of the accuracy of the estimate $\psi_{LRSE}(x_E, x_R)$. It is easily deduced from (7), and as discussed in [10], that there is a relationship between relative belief regions and the strength of the evidence (2). This makes sense as both are measuring the accuracy or reliability of inferences based on the measure of evidence as given by the relative belief ratio.

We investigated the choice of many other priors. The results were fairly consistent in obtaining evidence against $\psi = 0$ and with similar strengths. The biases did vary considerably with sometimes there being very low bias against $\psi = 0$ but accompanied by high bias in favor or conversely. Interestingly, such extreme cases are often accompanied by prior-data conflict.

Alternatively, we could assess the hypothesis of noninferiority, namely, $H_0 : \psi \in (-\delta, \infty)$ for which $\Pi((-\delta, \infty)) = 0.58$ and $\Pi((-\delta, \infty) | x_E, x_R) = 0.89$. Therefore,

$$RB_{\Psi}((-\delta, \infty)) = \frac{\Pi((-\delta, \infty) | x_E, x_R)}{\Pi((-\delta, \infty))} = \frac{0.89}{0.58} = 1.53$$

indicates evidence in favor of H_0 being true. Since the posterior probability of H_0 is large, this indicates strong evidence in favor of the experimental treatment being at least as effective as the reference treatment.

5 Conclusions

We have considered the application of relative belief inferences to an important inference problem with two-arm clinical trials. This is seen to provide a clear definition of what the evidence is, whether in favor of or against, for a hypothesis. Moreover, with such a definition this allows us to assess the bias introduced into a statistical analysis by a proper prior and so addresses a key concern with the use of Bayesian inference methods. In addition we have provided a methodology for eliciting an appropriate prior in such a context and demonstrated how one checks this prior to see if it is contradicted by the data.

In general, we take the view that all of statistics is subjective as we choose sampling models and priors and possibly even choose other ingredients. Such subjectivity is always present in a statistical analysis. Rather than searching for methods that are supposedly ‘objective’ in some sense, we embrace the subjectivity as allowing us to make judgements that reflect additional information we have about the application. Once the model and prior are chosen, we can go forward and make inference, based on the measure of statistical evidence and its calibration, in an unambiguous way. If we want these inferences to be convincing, however, it is important that we check the ingredients chosen against that aspect of the analysis that can best be claimed to be objective, namely, the data. So model checking and checking for prior-data conflict are essential parts of any statistical analysis. The analysis in this paper is presented as a meaningful application of this approach to statistical analyses.

6 Appendix

In Section 3.3 the prior predictive density of $V = (n_E + n_R - 2)s^2$ is

$$\begin{aligned} m_V(v) &= \int_0^\infty \frac{(v/\sigma^2)^{(n_E+n_R-2)/2-1} \exp\{-v/2\sigma^2\}}{2^{(n_E+n_R-2)/2} \Gamma((n_E+n_R-2)/2)} \frac{(1/\sigma^2)^{\alpha_0}}{\beta_0^{\alpha_0} \Gamma(\alpha_0)} \times \\ &\quad \exp\{-\beta_0/\sigma^2\} d(1/\sigma^2) \\ &= \frac{\Gamma((n_E+n_R-2+2\alpha_0)/2)}{\Gamma(\alpha_0)\Gamma((n_E+n_R-2)/2)} \left(\frac{v}{2\beta_0}\right)^{(n_E+n_R-2)/2-1} \times \\ &\quad \left(1 + \frac{v}{2\beta_0}\right)^{-(n_E+n_R-2)/2-\alpha_0} \frac{1}{2\beta_0}. \end{aligned}$$

Suppose $U = (U_1, U_2) | \sigma^2 \sim N_2(\mu, \sigma^2 \Sigma)$ where σ^2 is distributed as in

(8). Then the marginal density of U is

$$\begin{aligned} m_U(u) &= \int_0^\infty (2\pi)^{-1} (\det \Sigma)^{-1/2} \exp\{-(u - \mu)' \Sigma^{-1} (u - \mu)/2\sigma^2\} \frac{(1/\sigma^2)^{\alpha_0}}{\beta_0^{\alpha_0} \Gamma(\alpha_0)} \times \\ &\quad \exp\{-\beta_0/\sigma^2\} d(1/\sigma^2) \\ &= \frac{\Gamma(\alpha_0 + 1)}{\Gamma^2(1/2) \Gamma(\alpha_0)} (\det \Sigma)^{-1/2} (1 + (u - \mu)' \Sigma^{-1} (u - \mu)/2\beta_0)^{-(\alpha_0+1)} \beta_0^{-1} \end{aligned}$$

which is the density of a $t_{2\alpha_0}(2, \mu, (\beta_0/\alpha_0)\Sigma)$ distribution.

7 References

1. Snapinn, S. M. Noninferiority trials. *Current Controlled Trials in Cardiovascular Medicine* 2000, 1, 19-21.
2. Wellek, S. Testing Statistical Hypotheses of Equivalence and Noninferiority. *Chapman & Hall/ CRC*, 2010.
3. Gamalo, M.A., Muthukumarana, S., Ghosh, P. and Tiwari, R. C. A generalized p-value approach for assessing noninferiority in a three-arm trial. *Statistical Methods in Medical Research* 2013, 22, 261-277.
4. Berger, R. L. and Hsu, J. C. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 1996, 4, 283-319.
5. Hauschke, D. and Hothorn, L. A. Letter to the editor. *Statistics in Medicine*, 2007, 26, 230-236.
6. Gamalo, M. A., Wu, R. and Tiwari, R. C. Bayesian approach to non-inferiority trials for proportions. *Journal of Biopharmaceutical Statistics*, 2011, 21, 902-919.
7. Gamalo, M. A., Wu, R. and Tiwari, R. C. Bayesian approach to non-inferiority trials for normal means. *Statistical Methods in Medical Research*, 2012, online doi: 10.1177/0962280212448723.
8. Evans, M. Bayesian inference procedures derived via the concept of relative surprise. *Communications in Statistics* 1997, 26, 1125-1143.
9. Evans, M., Guttman, I., and Swartz, T. Optimality and computations for relative surprise inferences. *Canadian Journal of Statistics* 2006, 34, 113-129.

10. Baskurt, Z . and Evans, M. Hypothesis assessment and inequalities for Bayes factors and relative belief ratios. *Bayesian Analysis* 2013, 8, 3, 569-590.
11. Jeffreys, H. Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophy Society*, 1935, 31, 203-222.
12. Jeffreys, H. Theory of Probability (3rd ed.). *Oxford University Press* 1961.
13. Kass, R. E. and Raftery, A. E. Bayes factors. *Journal of the American Statistical Association*, 1995 90: 773-795.
14. Dickey, J. M. The weighted likelihood ratio, linear hypotheses on normal location parameters. *Annals of Statistics*, 1971 42: 204-223.
15. Evans, M. and Moshonov, H. Checking for prior-data conflict. *Bayesian Analysis*, 1, 4, 2006, 893-914.
16. Evans, M. and Jang, G. H. Invariant P-values for Model Checking. *Annals of Statistics*, 2010, 38, 1, 512-525.
17. Evans, M. and Jang, G. H. Weak informativity and the information in one prior relative to another. *Statistical Science*, 2011, 26, 3, 423-439.
18. Evans, M. and Jang, G. H. A limit result for the prior predictive. *Statistics and Probability Letters*, 2011, 81, 1034-1038.